The dataset consists the profit statement for 50 companies and various additional variables associated with those companies. This script is meant to develop a model where we can predict as to which particular variable(s) contribute most towards the profit of those companies and which company is most profitable

1. Import the dataset

```
dataset = read.csv('50_Startups.csv')
```

2. Encode categorical data

```
dataset$State = factor(dataset$State,
                       levels = c('New York', 'California', 'Florida'),
                       labels = c(1, 2, 3))
```

3. Split the dataset into training and test set

```
library(caTools)
set.seed(123)
split = sample.split(dataset$Profit, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
```

4. Fitting Multiple Linear Regression to the Training set

```
regressor = lm(formula = Profit ~ .,
               data = training_set)
```

5. Predict the test set results

```
y_pred = predict(regressor, newdata = test_set)
y_pred
```

```
##         4         5         8        11        16        20        21        24
## 173981.09 172655.64 160250.02 135513.90 146059.36 114151.03 117081.62 110671.31
##        31        32
##  98975.29  96867.03
```

```
test_set
```

```
##    R.D.Spend Administration Marketing.Spend State    Profit
## 4  144372.41      118671.85       383199.62     1 182901.99
## 5  142107.34       91391.77       366168.42     3 166187.94
## 8  130298.13      145530.06       323876.68     3 155752.60
## 11 101913.08      110594.11       229160.95     3 146121.95
## 16 114523.61      122616.84       261776.23     1 129917.04
## 20  86419.70      153514.11            0.00     1 122776.86
## 21  76253.86      113867.30       298664.47     2 118474.03
## 24  67532.53      105751.03       304768.73     3 108733.99
## 31  61994.48      115641.28        91131.24     3  99937.59
## 32  61136.38      152701.92        88218.23     1  97483.56
```

6. Result

As one can see, y_pred represents the model's predicted profiit values for the test set. The values are close to the test set.